

CIÊNCIA DE DADOS E INTELIGÊNCIA ARTIFICIAL

UFPB/CI/DCC

PROF. GUSTAVO OLIVEIRA

[GUSTAVO.OLIVEIRA@CI.UFPB.BR](mailto:GUSTAVO.OLIVEIRA@CI.UFPB.BR) | [GCPEIXOTO.GITHUB.IO](https://GCPEIXOTO.GITHUB.IO)

---

# INTRODUCÃO À CIÊNCIA DE DADOS

---

# OUTLINE

- ▶ História da Ciência de Dados
- ▶ Compreensões
- ▶ Perfis da área
- ▶ Ponto de vista acadêmico-científico x profissional
- ▶ Ciência de Dados no século XXI
- ▶ Propositura de temas de projeto

# HISTÓRIA DA CIÊNCIA DE DADOS

*"(...) a comprehensive and in-depth understanding of what data science is, and what can be achieved with data science and analytics research, education, and economy, **has yet to be commonly agreed.**"*

Cao L., 2017 (doi:10.1145/3076253)

# CRONOLOGIA

John W. Tukey  
(1915 - 2000)



["The future of data analysis"](#)

1962

"A ciência de lidar com dados, uma vez estabelecidos, enquanto a relação dos dados com o que representam é delegada a outros campos e ciências."  
(in: *Concise Survey of Computer Methods*)

**Missão:** "(...) converter dados em informação e conhecimento"

Fundação da IASC -  
[The International Association for Statistical Computing](#)

1977

1974

Peter Naur  
(1928 - 2016)



1º Workshop sobre  
Knowledge Discovery  
in Databases  
(Gregory Piatetsky-Shapiro)

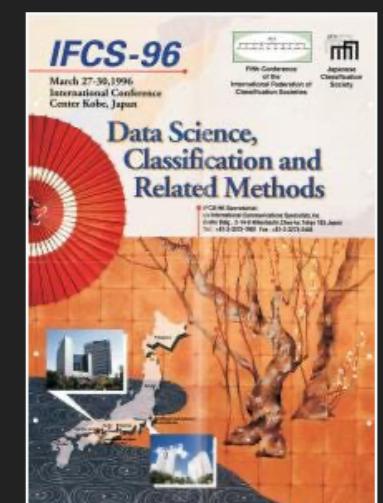
1989

**KDD [Knowledge Discovery in Databases]**  
refers to the overall process of discovering useful knowledge from data... **Data mining** is the application of specific algorithms for extracting patterns from data...

1999

1996

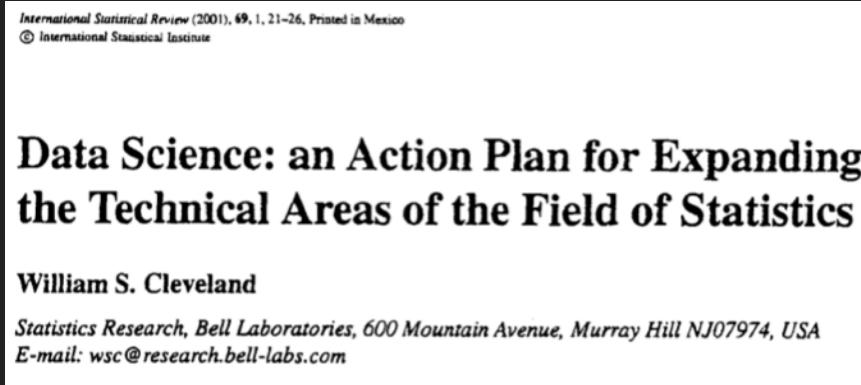
Conferência da *International Federation of Classification Societies (IFCS)* em Kobe, Japão, usa pela 1a. vez o termo *data science* em seu título



AI Magazine Volume 17 Number 3 (1996) (© AAAI)

**From Data Mining to Knowledge Discovery in Databases**

Usama Fayyad, Gregory Piatetsky-Shapiro, and Padhraic Smyth



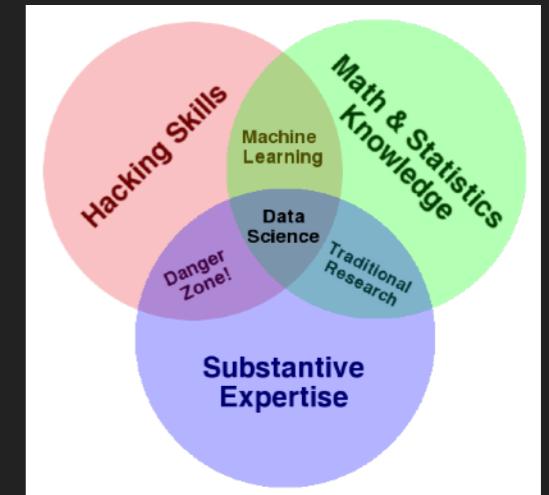
William Cleveland propôs um plano ambicioso para alargar o campo da estatística, que deveria ser chamado de **ciência de dados**

2001

Rise of the Data Scientist:  
*"We're seeing data scientists (...) emerge from the rest of the pack."*

Nathan Yau (aparentemente) cria o termo "cientista de dado"

2009



Drew Conway esboça o [Data Science Venn Diagram](#)

2010  
(SET)

2002  
Lançamento do *Data Science Journal* (CODATA)



2003  
Lançamento do *Journal of Data Science* (CODATA)

*"By 'Data Science' we mean almost everything that has something to do with data (...)"*

2010  
(JUN)

Mike Loukides publica o livro *What is Data Science?*



Tom Davenport e D. J. Patil publicaram o artigo Data Scientist: *The Sexiest Job of the 21st century* no Harvard Business Review

2012

*Top 50 Data Science Resources*  
por Molly Galetto

JOURNAL OF COMPUTATIONAL AND GRAPHICAL STATISTICS  
2017, VOL. 26, NO. 4, 745–766  
<https://doi.org/10.1080/10618600.2017.1384734>

## 50 Years of Data Science

David Donoho

Department of Statistics, Stanford University, Standford, CA

2017  
(AGO)

2017  
(JUN)

**Data Science: A Comprehensive Overview**  
LONGBING CAO, University of Technology Sydney, Australia

Longbin Cao publica o artigo *Data Science: Profession and Education* na IEEE Computer Society

*"However, despite the fact that the role of data scientists has been described as the sexiest job in the 21st century, the qualifications and capabilities of a data scientist are not clearly defined; it is important yet undermined to define what makes the next-generation data scientists who can transform today's and future science, technology, innovation, and economy"*

2019

2018

Kees Groeneveld publica o artigo *The New Sexiest Job of the 21st* no LinkedIn

Turma do CI inicia  
curso de ICD  
na UFPB

2021

VOCÊ DIRÁ  
O RESTO  
DA HISTÓRIA!



# COMPREENSÕES

## DATA X DATUM

- ▶ "*Datum*", no inglês, equivale a "dado", no singular
- ▶ "*Data*", no inglês, equivale a "dados", no plural
- ▶ No uso corrente, "*data*" tem sido usado tanto no plural quanto no singular
- ▶ "Dados [*data*] são fatos, estatísticas ou itens de informação individuais, geralmente numéricos, que são coletados por meio da observação. Em um sentido mais técnico, os dados [*data*] são um conjunto de valores de variáveis qualitativas ou quantitativas sobre uma ou mais pessoas ou objetos, enquanto um dado [*datum*] (singular de dados) é um valor único de uma única variável." (Wikipedia)
- ▶ Obs.: o termo "datum" utilizado em georeferenciamento e cartografia não deve ser confundido com o "dado" no sentido computacional.



## DEFINIÇÕES DE "DATA SCIENCE" A PARTIR DE MÚLTIPLAS PERSPECTIVAS

- ▶ **(Perspectiva de alto-nível)**

*"Ciência de Dados é a ciência dos dados ou o estudo dos dados"*

- ▶ **(Perspectiva disciplinar)**

*"Um novo campo interdisciplinar que sintetiza e se constrói sobre a estatística, informática, computação, comunicação, gestão e sociologia para estudar dados e seus ambientes (incluindo domínios e outros aspectos contextuais, tais como sociais e organizacionais) a fim de transformar dados em insights e decisões seguindo um pensamento e uma metodologia dado > conhecimento > sabedoria"*

## CONT.

- ▶ *"Ciência de Dados é a extração do conhecimento útil diretamente a partir de dados através de um processo de descoberta ou de formulação e teste de hipóteses."*  
*(NIST 1500-1, EUA, 2015)*
- ▶ *"Ciência de Dados é um campo de estudo e prática que envolve a coleta, o armazenamento e o processamento de dados para obter percepções importantes sobre um problema ou fenômeno."* *(Chirag Shah)*

---

## CONT.

- ▶ "Ciência de Dados é a extração do conhecimento útil diretamente a partir de dados através de um processo de descoberta ou de formulação e teste de hipóteses."  
(NIST 1500-1, EUA, 2015)
- ▶ "Ciência de Dados é um campo de estudo e prática que envolve a coleta, o armazenamento e o processamento de dados para obter percepções importantes sobre um problema ou fenômeno." (Chirag Shah)
- ▶ A novidade da ciência de dados não está enraizada no conhecimento científico mais recente, mas em uma **mudança disruptiva em nossa sociedade que foi causada pela evolução da tecnologia**: a datificação. A datificação é o processo de renderização em aspectos de dados do mundo que nunca foram quantificados antes." (Laura Igual & Santi Seguí)

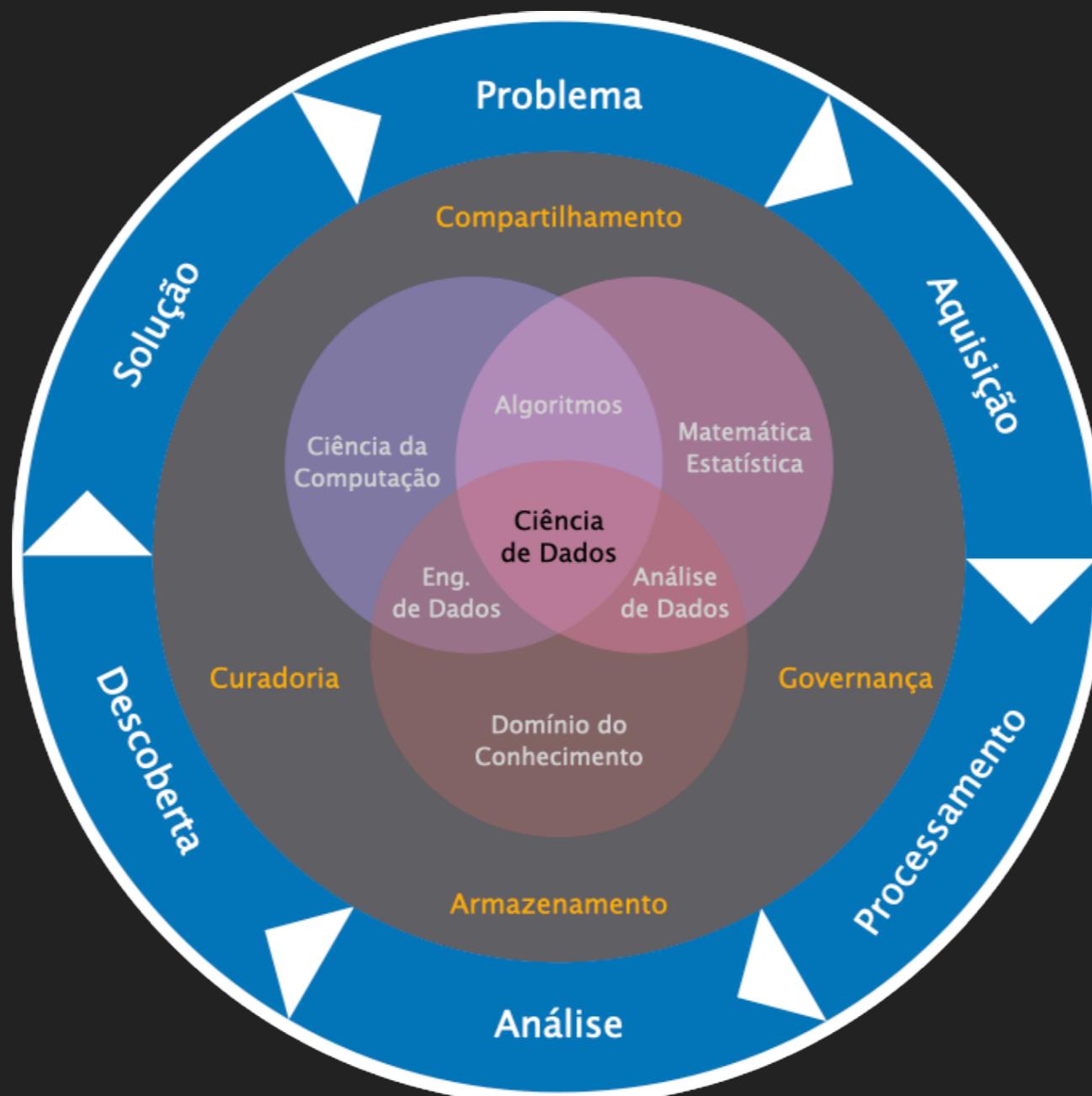
## O QUE DIZ O INTERNATIONAL JOURNAL OF DATA SCIENCE AND ANALYTICS

- ▶ "Data Science has been established as an important **emergent scientific field and paradigm driving research evolution** in such disciplines as statistics, computing science and intelligence science, and **practical transformation** in such domains as science, engineering, the public sector, business, social science, and lifestyle. The field encompasses the larger areas of artificial intelligence, data analytics, machine learning, pattern recognition, natural language understanding, and big data manipulation. It also **tackles related new scientific challenges**, ranging from data capture, creation, storage, retrieval, sharing, analysis, optimization, and visualization, to integrative analysis across heterogeneous and interdependent complex resources **for better decision-making, collaboration, and, ultimately, value creation.**"

## PRODUTOS DE DADOS

- ▶ Um produto de dados é algo habilitado ou orientado por dados e pode ser uma descoberta, previsão, serviço, recomendação, insight de tomada de decisão, pensamento, modelo, modo, paradigma, ferramenta ou sistema. Os produtos de dados de valor são **conhecimento, inteligência, sabedoria e decisão.**

## VISÃO HOLÍSTICA DA CIÊNCIA DE DADOS



► **Matemática/Estatística**

- modelos matemático; análise e inferência de dados; aprendizagem de máquina

► **Ciência da Computação/Engenharia de Software**

- hardware/software; projeto, armazenamento e segurança de dados

► **Conhecimento do Domínio/Expertise**

- ramo de aplicação do conhecimento; *data reporting* e inteligência de negócios; marketing e comunicação de dados

- ▶ 1. **Definição do problema**, etapa em que uma "grande pergunta" é feita, a qual, a princípio, pode ser respondida ao se vasculhar um conjunto de dados específico.
- ▶ 2. **Aquisição de dados**, etapa em que se coleta toda a informação relacionada ao problema lançado na etapa anterior.
- ▶ 3. **Processamento de dados**, etapa em que os dados adquiridos são processados para análise. Nesta etapa realiza-se um verdadeiro tratamento dos dados (limpeza, formatação e organização).
- ▶ 4. **Análise de dados**, etapa em que os dados são analisados e perscrutados por meio de técnicas de mineração, agrupamento e clusterização. Neste momento é que testes de hipótese e mecanismos de inferência são utilizados.
- ▶ 5. **Descoberta de dados**, etapa em que descobertas são realizadas, tais como correlações entre variáveis, comportamentos distintivos e tendências claramente identificáveis, permitindo que conhecimento seja gerado a partir da informação.
- ▶ 6. **Solução**, etapa final do ciclo na qual as descobertas podem ser convertidas em produtos e ativos de valor agregado para o domínio do problema proposto.

## 4 ESTRATÉGIAS PARA EXPLORAR O MUNDO USANDO DADOS

- ▶ **Fazer uma sondagem da realidade:** só decidiremos por A ou B depois de examinar A e B.
- ▶ **Descobrir padrões:** problemas "datificados" são analisáveis por algoritmos, modelos matemáticas e técnicas de programação
- ▶ **Predizer eventos futuros:** não se pode prever o futuro, mas identificar eventos previsíveis em algum sentido representa um conhecimento altamente valioso
- ▶ **Entender as pessoas e o mundo:** é um objetivo audacioso e de difícil conclusão, mas decisões otimizadas podem ser feitas com o uso de modelos comportamentais, aprendizagem profunda e processamento de linguagem.

# CIÊNCIA DA COMPUTAÇÃO X CIÊNCIA DE DADOS X CIÊNCIA REAL

- ▶ O problema é o "mindset"...
- ▶ Você age como você pensa
- ▶ **Dados x "método-centrismo":**  
Cientistas são guiados por dados  
Cientistas da computação são guiados por algoritmos e  
"obcecados" pelo método (que algoritmo é melhor?  
Python, Java, C++?)  
Cientistas reais gastam tempo coletando dados para  
responder suas perguntas



## CONT.

- ▶ **Preocupação com resultados:**

Cientistas de verdade preocupam-se com respostas

Bons cientistas analisam a descoberta e preocupam-se  
com o que significam

Maus cientistas da computação preocupam-se em  
produzir números plausíveis

## CONT.

### ► Precisão:

Na ciência de verdade, nada é completamente verdadeiro ou falso

Na ciência da computação ou matemática, quase sempre algo é verdadeiro ou falso

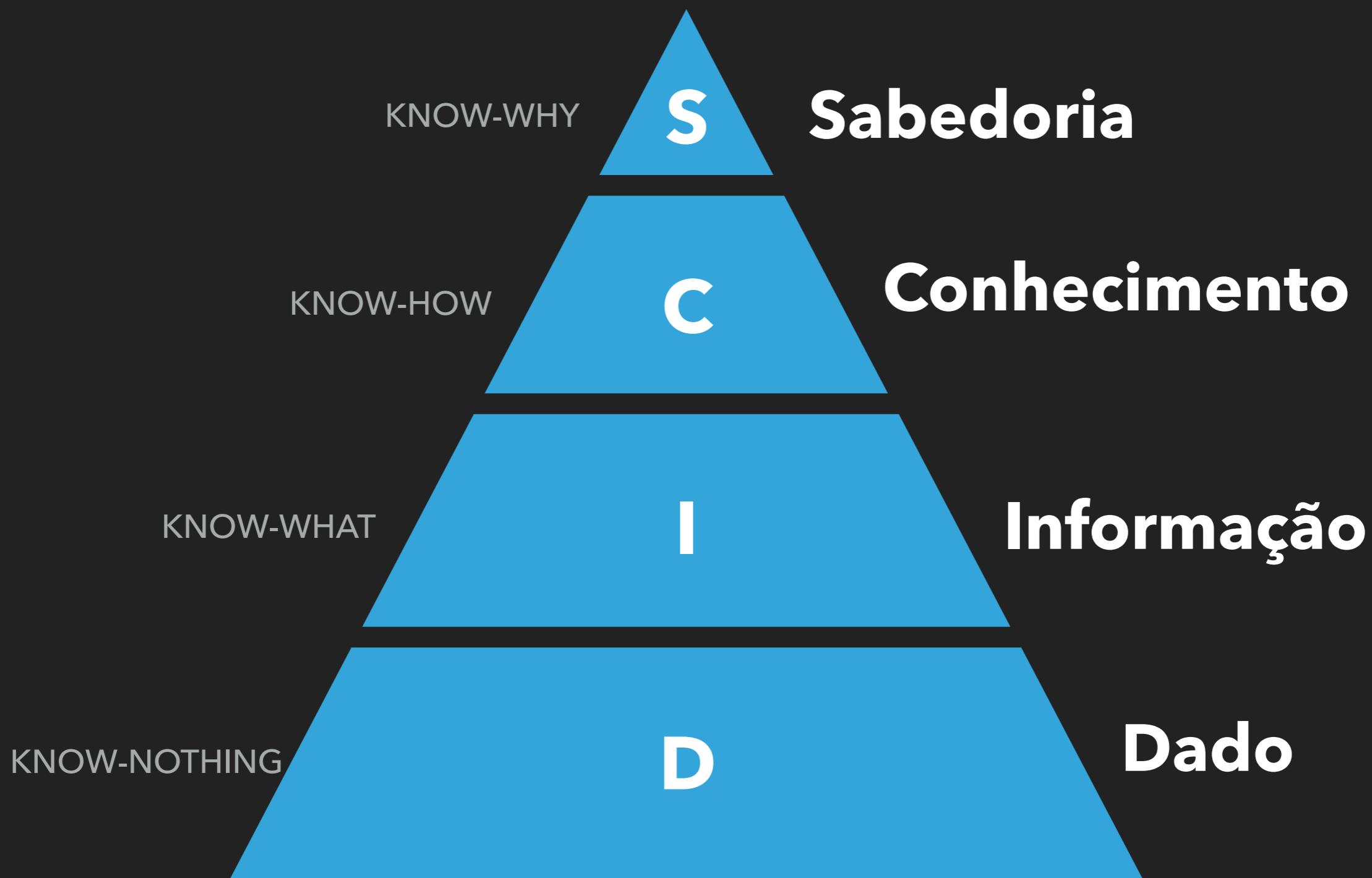
Cientistas da computação gostam de números em ponto flutuante

Cientistas de verdade preocupam-se com o que os números "dizem"

## ASPIRANTES A CIENTISTAS DE DADOS DEVEM...

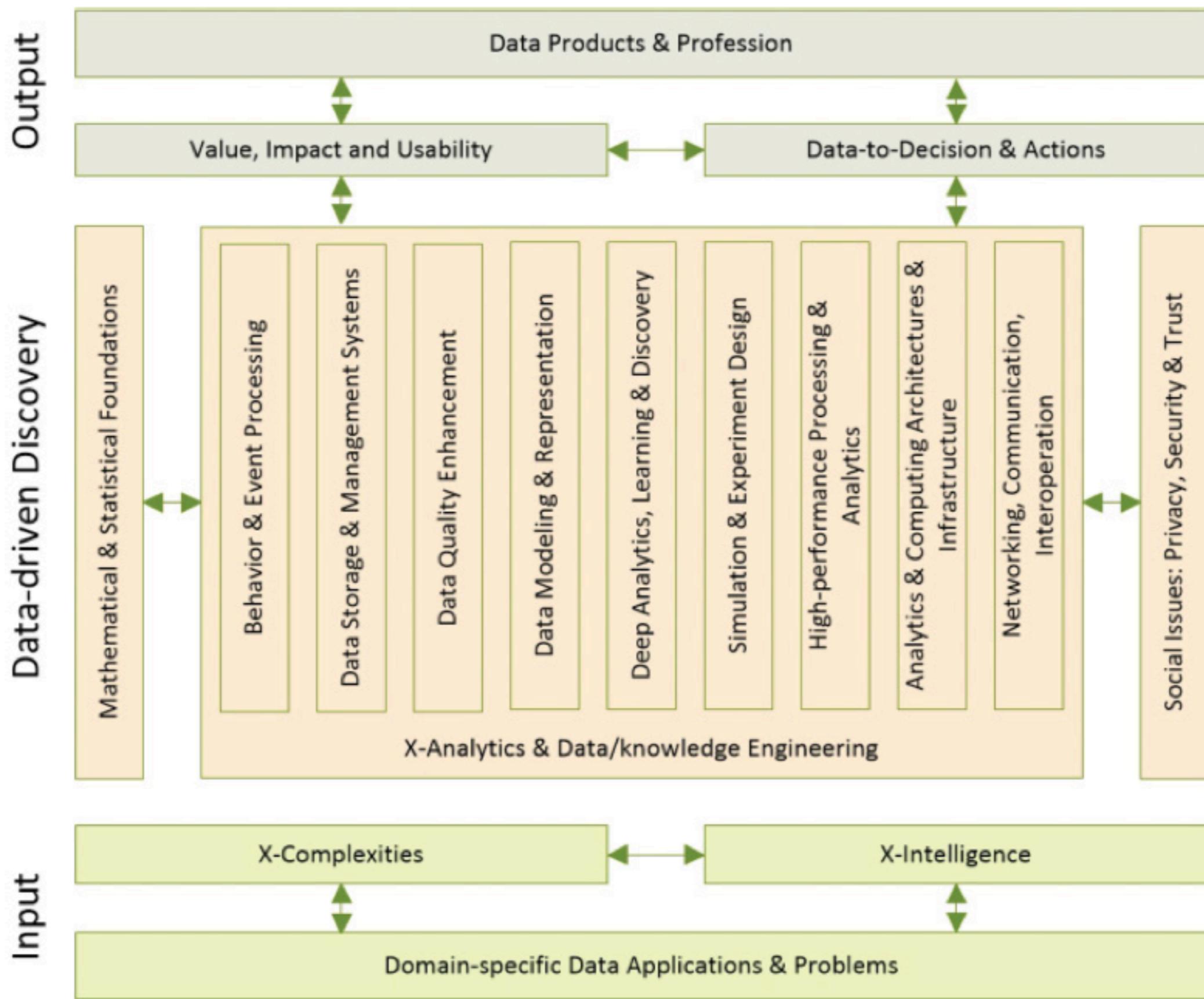
- ▶ Pensar como cientistas de verdade
- ▶ Converter números em insights
- ▶ Entender os "porquês" e os "comos"
- ▶ Ter um amplo espectro de interesses
- ▶ Ser corajosos para deixar a zona de conforto
- ▶ Aprender a fazer perguntas interessantes para os dados

## A PIRÂMIDE DICS (DIKW)



# MAPA CONCEITUAL DA CIÊNCIA DE DADOS

25



# PERFIS DA ÁREA

## CIENTISTA DE DADOS X ANALISTA DE DADOS X ENGENHEIRO DE DADOS

- ▶ **Cientista de dados** é um profissional que tem conhecimentos suficientes sobre necessidades de negócio, domínio do conhecimento, além de possuir habilidades analíticas, de software e de engenharia de sistemas para gerir, de ponta a ponta, os processos envolvidos no ciclo de vida dos dados.

## CONT.

- ▶ **Analista de dados** é o profissional capaz de sintetizar conhecimento a partir da informação e convertê-lo em ativos exploráveis.
- ▶ **Engenheiro(a) de dados** é o(a) profissional que explora recursos independentes para construir sistemas escaláveis capazes de armazenar, manipular e analisar dados com eficiência e desenvolver novas arquiteturas sempre que a natureza do banco de dados exigi-las.

# MATRIZ DE COMPETÊNCIAS

CD	AD	ED
<ul style="list-style-type: none"><li>• Realiza o pré-processamento, a transformação e a limpeza dos dados;</li><li>• Usa ferramentas de aprendizagem de máquina para descobrir padrões nos dados;</li><li>• Aperfeiçoa e otimiza algoritmos de aprendizagem de máquina;</li><li>• Formula questões de pesquisa com base em requisitos do domínio do conhecimento;</li></ul>	<ul style="list-style-type: none"><li>• Analisa dados por meio de estatística descritiva;</li><li>• Usa linguagens de consulta a banco de dados para recuperar e manipular a informação;</li><li>• Confecciona relatórios usando visualização de dados;</li><li>• Participa do processo de entendimento de negócios;</li></ul>	<ul style="list-style-type: none"><li>• Desenvolve, constroi e mantém arquiteturas de dados;</li><li>• Realiza testes de larga escala em plataformas de dados;</li><li>• Manipula dados brutos e não estruturados;</li><li>• Desenvolve <i>pipelines</i> para modelagem, mineração e produção de dados;</li><li>• Cuida do suporte a cientistas e analistas de dados;</li></ul>

CDIA

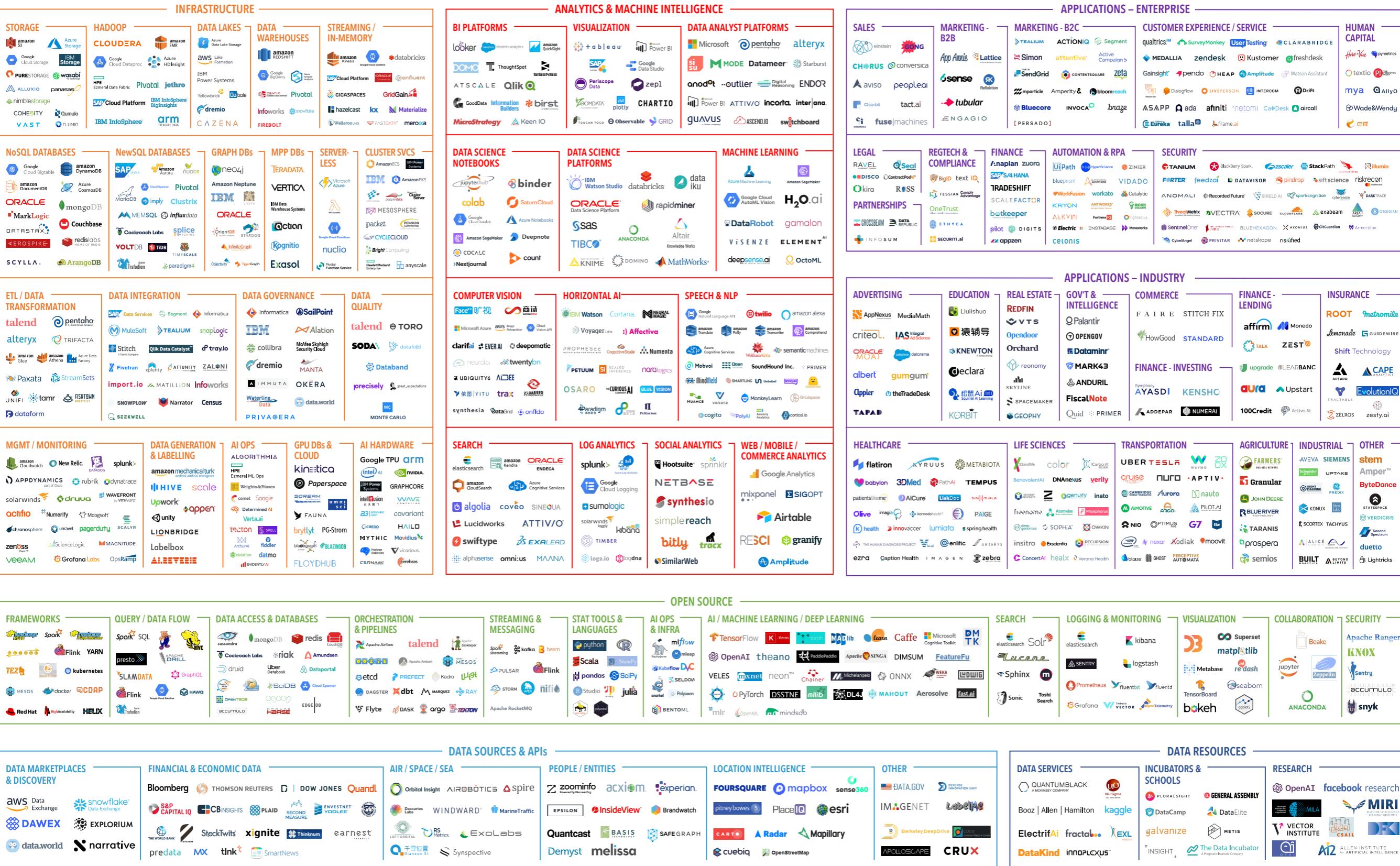
---

LANDSCAPE  
2020

# FERRAMENTAS, PLATAFORMAS, EMPRESAS E APLICAÇÕES

31

## DATA & AI LANDSCAPE 2020



# PONTO DE VISTA ACADÊMICO-CIENTÍFICO X PROFISSIONAL

# O QUE A UNIVERSIDADE PROPÕE

- ▶ O escopo teórico faz parte do todo
- ▶ Nem tudo pode ser "agradável", à primeira vista
- ▶ Importa que você saiba "pensar *data science*" e não apenas ser um profissional cego aplicador de *data science*
- ▶ Em alguns anos, você também pode se tornar um grande pesquisador

## O QUE O MERCADO/SETOR EXTERNO PROPÕE

- ▶ Experiências com situações reais
- ▶ Necessidade imediata de entregas
- ▶ Menor tempo com pesquisa
- ▶ Geração de valor

# COMPETÊNCIAS FORMATIVAS

Pensamento  
Bases  
Fundamentos  
Cognição  
Matemática  
Estatística  
Otimização  
Análise  
Representação  
Modelagem

TÉCNICAS  
DE  
ENGENHARIA

PENSAR  
SOBRE  
CD

BASES  
TEÓRICAS

PRÁTICAS

COMUNICAÇÃO

GESTÃO

LIDERANÇA

VOCÊ  
EM ICD

VOCÊ ESTUDANTE  
NO RESTANTE DO CURSO

VOCÊ  
PROFISSIONAL

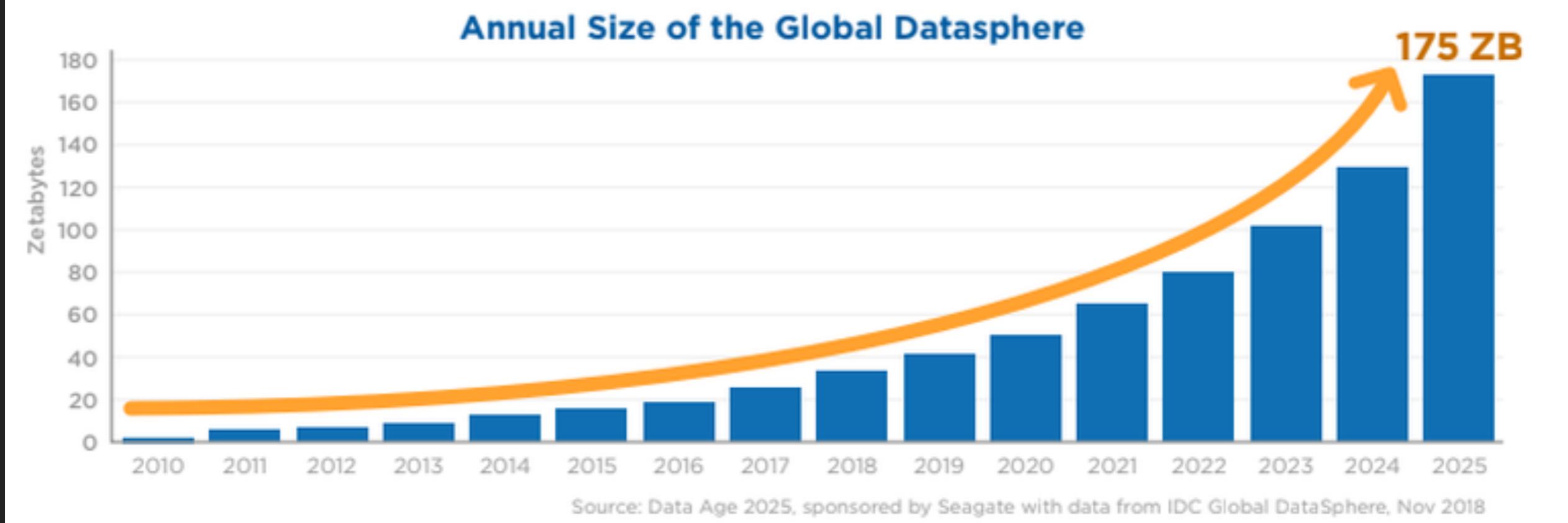
DIFUSÃO MULTISETORIAL

---

CIÊNCIA DE DADOS NO  
SÉCULO XXI

# A DATASFERA

Figure 1 – Annual Size of the Global Datasphere



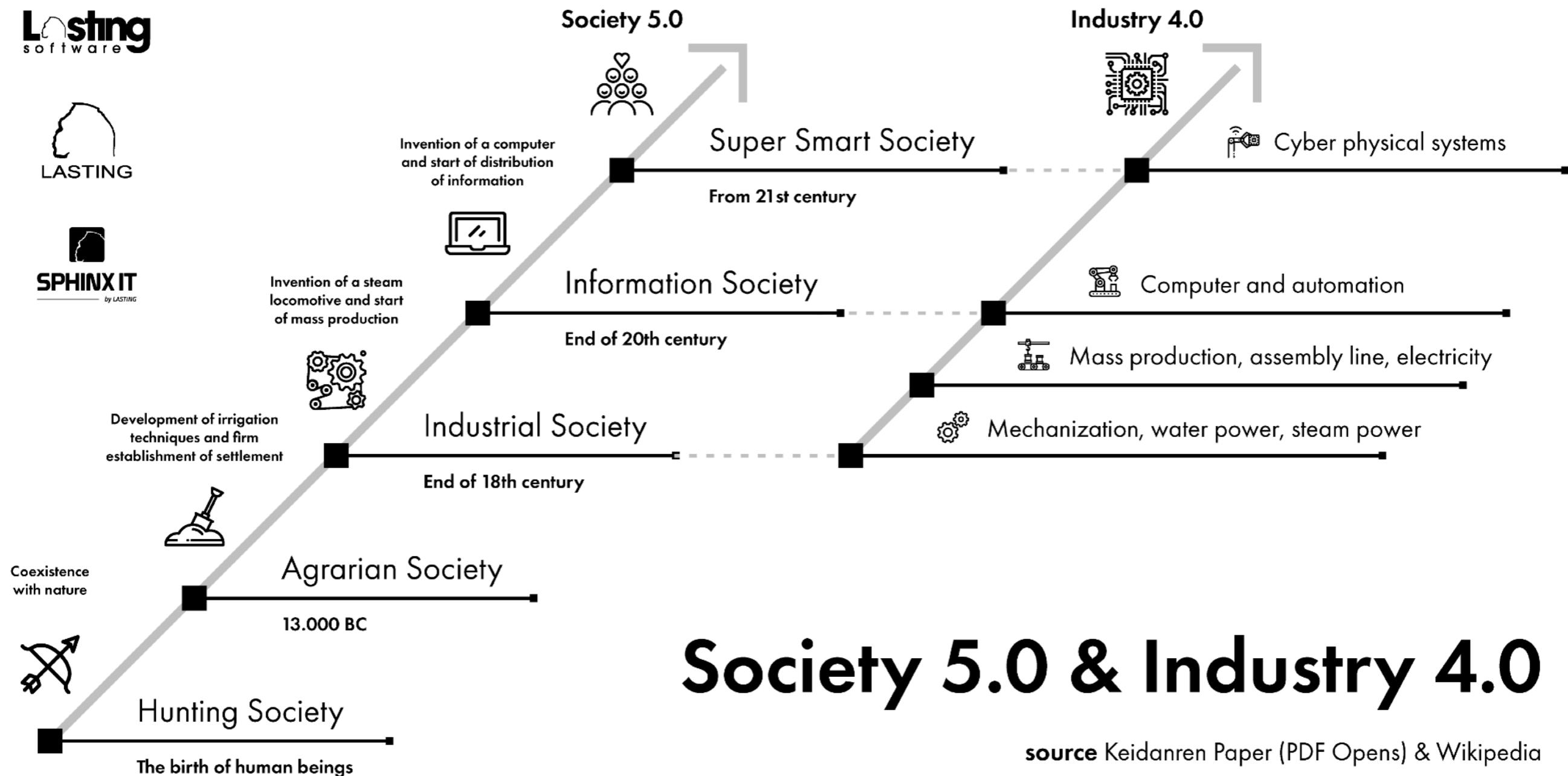
1 ZB = 1e9 TB. Levando em consideração que o mundo possua 8 bi de pessoas em 2025, a quantidade gerada de dados per capita naquele ano será de aproximadamente 22 TB = 22 HDs de 1 TB por pessoa!

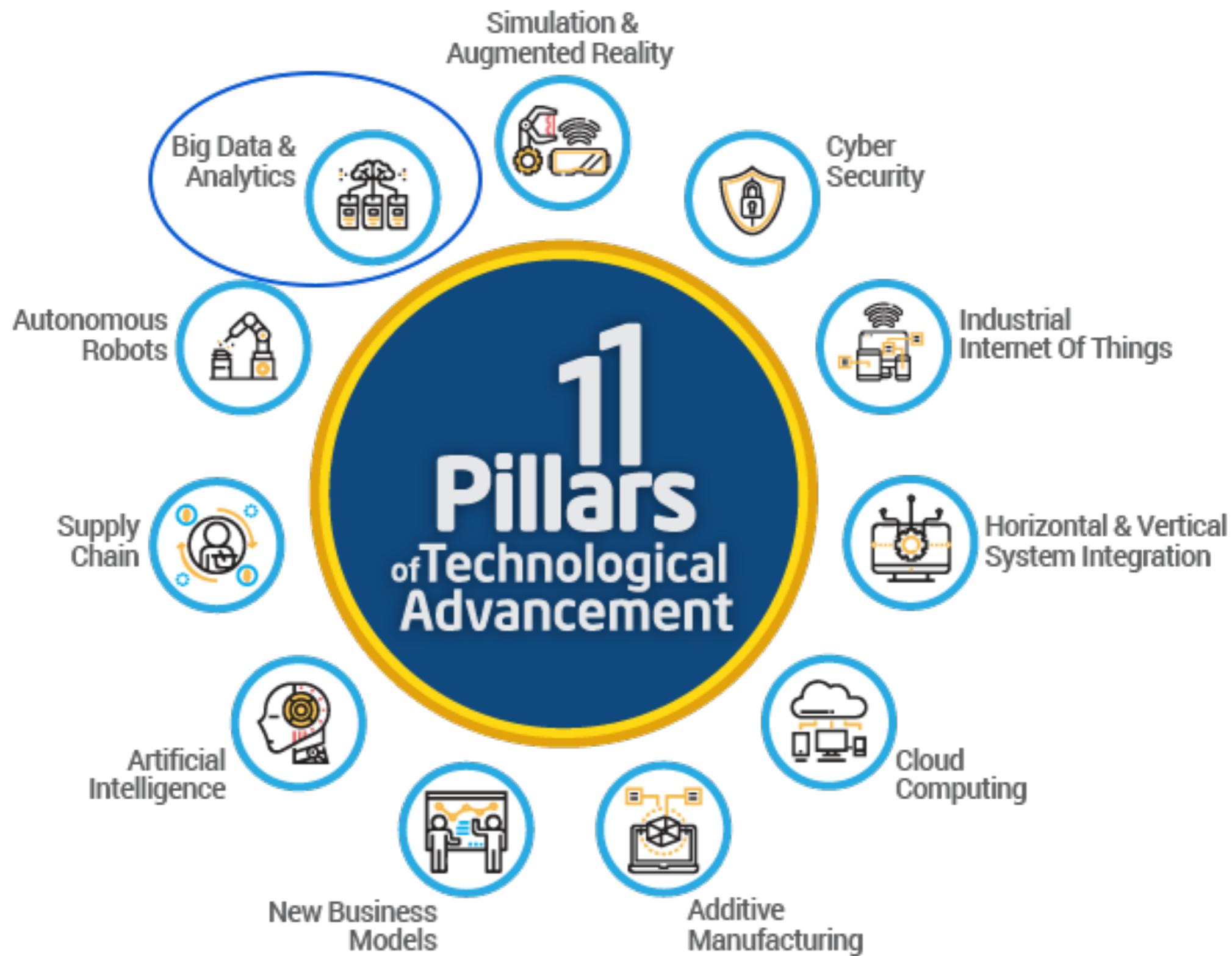
# INDÚSTRIA 4.0 / SOCIEDADE 5.0

**Lasting**  
software



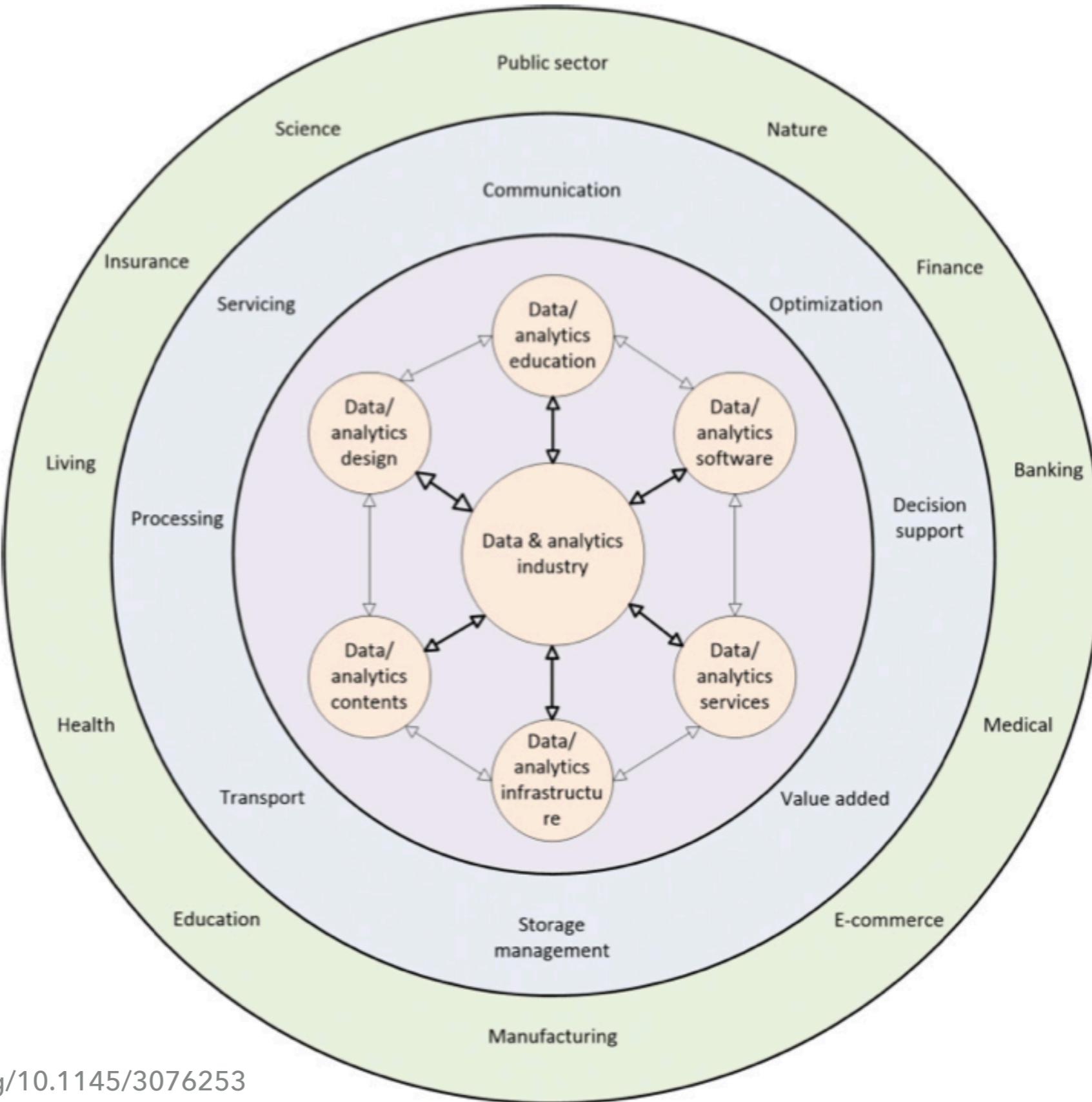
**SPHINX IT**  
by LASTING





# DIFUSÃO DA CIÊNCIA DE DADOS E ANALYTICS

40



# EXEMPLOS

- ▶ FINANCEIRO
  - ▶ Modelos preditivos de eventos de mercado
  - ▶ Detecção de fraudes bancárias
  - ▶ Mitigação de riscos em operações em bolsa
  - ▶ Análise de crédito de clientes

# CONT.

- ▶ POLÍTICAS PÚBLICAS
  - ▶ Planos de ação para melhorias de transporte público
  - ▶ Insights sobre tráfego, bem estar e saúde pública
  - ▶ Mapeamento de nível de contágio (e.g. COVID-19)

# CONT.

- ▶ POLÍTICA
  - ▶ Melhoria de modelos de votação e participação popular
  - ▶ Captação de eleitores em mídias sociais
  - ▶ Verificação de cumprimentos de campanha
  - ▶ Transparência

# CONT.

- ▶ SAÚDE
  - ▶ Sequenciamento genético
  - ▶ Diagnóstico precoce de doenças graves
  - ▶ Monitoramento de sinais vitais
  - ▶ Gestão hospitalar

## CONT.

- ▶ PLANEJAMENTO URBANO
  - ▶ *Smart cities*
  - ▶ Projetos de infraestrutura
  - ▶ Ações em vizinhanças
  - ▶ Gestão hídrica e saneamento

# CONT.

- ▶ EDUCAÇÃO
  - ▶ Melhorias de ambientes de aprendizagem
  - ▶ Descoberta de meios preferenciais de aprendizagem
  - ▶ Análise de desempenho



PROJETOS CAPSTONE

---

PROPOSITURA  
DE TEMAS

- ▶ Artes, entretenimento (cinema, teatro, música)
- ▶ Clima, meio ambiente, recursos naturais
- ▶ Saúde, esportes, qualidade de vida
- ▶ Agricultura, pecuária, pesca
- ▶ Energias, transição energética
- ▶ Indústria 4.0, transformação digital, IoT
- ▶ Economia, renda, desigualdade social, finanças, mercados
- ▶ Direito, legislação, justiça, segurança pública, computação forense
- ▶ Política, ética, cidadania
- ▶ Infraestrutura, cidades inteligentes
- ▶ Turismo, lazer
- ▶ Geopolítica, relações internacionais, comércio exterior
- ▶ Biotecnologia, materiais avançados
- ▶ Gestão, administração, pessoas
- ▶ Educação, ciência, tecnologia, inovação

DICA 1:

JÁ COMECE A PENSAR  
NO QUE VOCÊ GOSTA E  
COMO PENSA EM CONTRIBUIR

DICA 2:

CONSULTE OS OBJETIVOS DE  
DESENVOLVIMENTO SUSTENTÁVEL  
DA AGENDA 2030 DA ONU

# A ÚLTIMA MENSAGEM

## O QUE ESPERAR DESTE CURSO

PENSAR SOBRE A CD (INCLUSIVE FILOSOFICAMENTE)  
SER INTRODUZIDO À CD  
COMPREENDER OS FUNDAMENTOS DA CD

## O QUE NÃO ESPERAR DESTE CURSO

SABER RESOLVER TODO TIPO DE PROBLEMA  
REALIZAR ANALYTICS PROFUNDO  
SER UM PROGRAMADOR "DA PESADA"

---

**BOM CURSO E SUCESSO!**